취업정보 동향 분석

Jinseog Kim
Dongguk University
jskim1986@gmail.com

2018-03-12

Contents

1 취업정보 분석배경 및 목적

- 최근 국내외 경제 여건 악화로 대졸 취업난 가중
- 정부의 대학 교육체계 전환 요구(국가직무능력표준(NCS)의 제정, PRIME사업)
- 대학 및 관련학과의 교육 프로그램 개선 시급
- 국내 채용사이트의 통계학 전공관련 채용정보 기초 분석
- 주요 키워드에 대하여 커뮤니티 탐색방법을 통해 키워드간 연관성을 고찰
- 학과졸업대상자의 취업 준비 및 교육 프로그램 개선을 위한 기초자료 제공

2 채용정보의 수집

- http://kr.indeed.com/
  - 채용정보사이트
  - 자체 + 타 채용사이트 정보포함
- 수집 방법 및 도구
  - R패키지 XML, Rcurl을 이용
  - 검색어 : 통계, 데이터분석, 빅데이터
  - 제목, 회사명, 지역, 공고내용을 수집

```r
library(XML)
library(RCurl)
getText <- function(i){
  url <- "http://kr.indeed.com/%EC%B7%A8%EC%97%85?q=%ED%86%B5%EA%B3%84"
  url0  <- "http://kr.indeed.com/%ED%86%B5%EA%B3%84%EC%A7%81-%EC%B7%A8%EC%97%85"
  url <- ifelse(i==0, url0, paste0(url, "&start=", i*10))#430
  doc <- getURL(url, .encoding="UTF-8")
  d <- htmlParse(doc)
  jobtitle <- xpathSApply(d, '//*[@class="jobtitle"]', xmlValue)
  company <- xpathSApply(d, '//*[@class="company"]', xmlValue)
  location <- xpathSApply(d, '//*[@class="location"]', xmlValue)
  description <- xpathSApply(d, '//*[@class="summary"]', xmlValue)
  list(jobtitle, company, location, description)
}
raw_doc <- lapply(1:290, getText)
```

3 데이터 처리

3.1 전처리

```
data_process <- function(x, i){
    require(stringr)
    x <- do.call(c, lapply(raw_doc, "[[", i))
    x <- gsub("[^A-Za-z가-힣[:space:][:digit:][:punct:]]", "", x)
    x <- gsub("\n", " ", x)
    x <- gsub("[[:digit:]]", " ", x)
    x <- gsub("[[:punct:]]", " ", x)
    x <- tolower(x)
    for(i in 1:5) x <- gsub("  ", " ", x)
    x <- str_trim(x)
    x
}
```

```
jobtitle <- data_process(raw_doc, 1)
company <- data_process(raw_doc, 2)
tab_c <- table(company)
tab_c[tab_c > 1]
```

3.2 지역명 프로세싱

```
location <- data_process(raw_doc, 3)
location <- strsplit(location, " ")
location <- do.call("c", lapply(location, function(x) x[1]))

i11 <- grep("서울", location)
i21 <- grep("부산", location)
i22 <- grep("대구", location)
i23 <- grep("인천", location)
i24 <- grep("광주", location)
i25 <- grep("대전|세종", location)
i26 <- grep("울산", location)
i31 <- grep("경기|고양|과천|구리|군포|남양주|부천|성남|수원|시흥|안산|안성|안양|여주|용인|의왕|의정
i32 <- grep("강원|원주|홍천", location)
i33 <- grep("충청북|충북|청주|충주|음성|진천|증평", location)
i34 <- grep("충청남|충남|아산|당진", location)
i35 <- grep("전라남|전남|무안|완주|순천", location)
i36 <- grep("전라북|전북|전주", location)
i37 <- grep("경북|경상북도|구미|성주|영주|영천|포항", location)
i38 <- grep("경남|경상남도|산청|양산|진주|창원", location)
```

```r
i39 <- grep("제주¦서귀", location)
location[i11] <- "서울"
location[i21] <- "부산"
location[i22] <- "대구"
location[i23] <- "인천"
location[i24] <- "광주"
location[i25] <- "대전/세종"
location[i31] <- "경기"
location[i32] <- "강원"
location[i33] <- "충북"
location[i34] <- "충남"
location[i35] <- "전북"
location[i36] <- "전남"
location[i37] <- "경북"
location[i38] <- "경남"
location[i39] <- "제주"
```

- table(location)

4 기초분석

```
t_job <- table(jobtitle)
t_com <- table(company)

T_all <- paste(jobtitle, company, location)
t_all <- table(T_all)
T_over_2 <- names(t_all)[t_all>1]
drop_idx <- lapply(1:length(T_over_2), function(x){
    which(T_all == T_over_2[x])[-1]
})
drop_idx <- do.call(c, drop_idx)


#[1:100] #945개
```

4.1 메타데이터

```
meta <- data.frame(jobtitle, location)[-drop_idx,]
```

4.2 채용공고 프로세싱

```
description <- data_process(raw_doc, 4)
description <- description[-drop_idx]
library(KoNLP)
library(tm)
useSejongDic()
```

5 데이터를 DTM으로 변환

5.1 명사 추출

```
jj2 <- lapply(1:length(description), function(x) {
    texts <- paste(meta$jobtitle[x], description[x])
  y <- extractNoun(texts)
  #y <- extractNoun(x1[1])
  y[sapply(y, nchar) > 1]
 }
);
```

5.2 unique terms 추출

```
uterms <- unique(unlist(jj2))
uterms <- sort(uterms)
length_terms <- sapply(uterms, function(x) nchar(x))
```

5.3 15자 이상인 단어 제거

```
uterms <- uterms[length_terms<15]
#uterms <- uterms[is.na(match(uterms, c("분석", "데이터", "통계")))]
```

5.4 문서-단어 행렬

```
# simple_triplet matrix form으로 변환
s_mat <- lapply(seq_along(jj2), function(i){
    x <- match(jj2[[i]], uterms)
    out <- table(x)
```

```
    i <- rep(i, length(out))
    j <- as.integer(names(out))
    v <- unname(out)
    cbind(i, j, v)
})

s_mat <- do.call("rbind", s_mat)
# DTM
s_mat_s <- slam::simple_triplet_matrix(s_mat[,1], s_mat[,2], s_mat[,3], nrow=length(jj2),
        dimnames = list(doc = as.character(seq_along(jj2)), terms=uterms))

class(s_mat_s) <- c("DocumentTermMatrix", "simple_triplet_matrix")
##
s_mat_bin <- weightBin(s_mat_s)
s_mat_tfidf <- weightTfIdf(s_mat_s)
```

5.5 R 데이터로 저장

```
save(raw_doc, s_mat_s, s_mat_bin, s_mat_tfidf, meta, file="recruit.RData")
```

5.6 희소빈도 단어 제거

```
dtm <- removeSparseTerms(s_mat_bin, sparse=0.95)
# n terms = 366 if docs occured terms < 4
#=======================================================
```

5.7 채용공고의 워드클라우드 분석

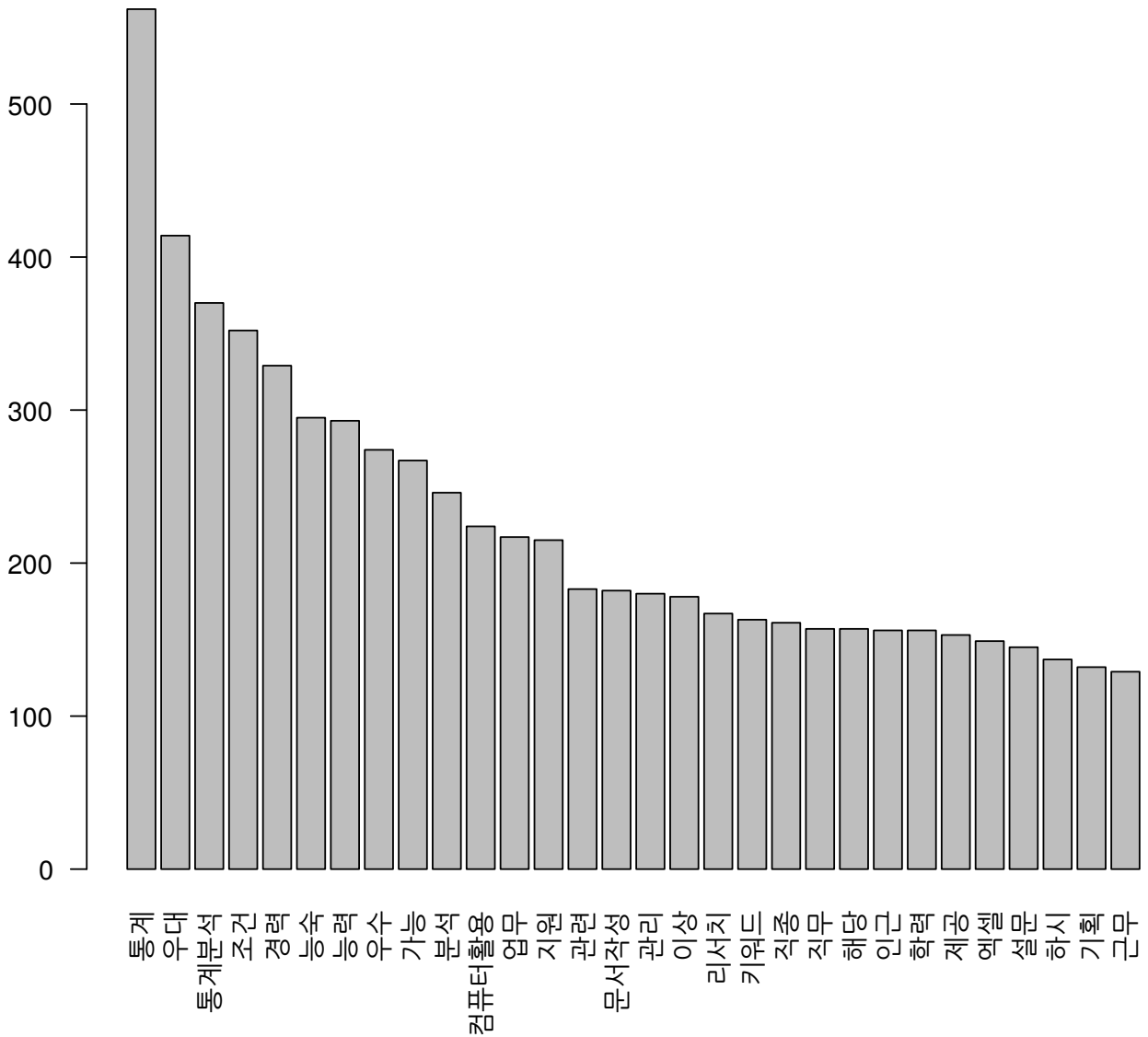
```
load("recruit.RData")
library(wordcloud)
library(slam)
pal <- brewer.pal(8,"Dark2")
wordFreq <- sort(col_sums(s_mat_bin), decreasing=TRUE)
wordcloud(words=names(wordFreq), freq=wordFreq,
          min.freq=50,
            random.order=F,
```

```
            random.color=T,
            colors=pal)
```

통계
우대
통계분석
조건
능숙
경력
컴퓨터활용
우수
분석
가능
문서작성
업무
지원
관리
리서치
이상
키워드
직무
직종
근무지역
학력
무관
인근
담당
인턴
제공
엑셀
하시
기획
관련
서울
현황
조사분석
회사
모집
활용
통계학
설문
데이터
근무
온라인
영어
이용
고급
개발
필요
마케팅
연봉
조사
연구
사무
근무경험
확인
지원자
운영
경험자
담당자
수정
작성
건강보험
로그인
회사명
커피
신입
거주
경영
졸업
회계
활동
대보험
성별
운전
참조
채용
입사
payopen
주소
사원
무료
직원
형태
소지자
공고
연구소
고객
자료
sas
업종
경력자
데이타
요청
보조
고용
정보
오픈
전공
컨설팅
증명서
사항
페이
온라
점검
차량
분야
경험

5.8 중요 키워드

```
par(mar=c(6,4,2,2))
barplot(wordFreq[1:30], las=2)
```


500
400
300
200
100
0
통계
우대
통계분석
조건
경력
능숙
능력
우수
가능
분석
컴퓨터활용
업무
지원
관련
문서작성
관리
이상
리서치
키워드
직종
직무
해당
인근
학력
제공
엑셀
설문
하시
기획
근무

5.9 키워드 그룹 탐색 및 연관네트워크

- sparse keyword제거

```
new.x <-  as.matrix(s_mat_bin)
x <- slam::crossprod_simple_triplet_matrix(s_mat_bin)#/diag(s_mat_bin)
library(igraph)
```

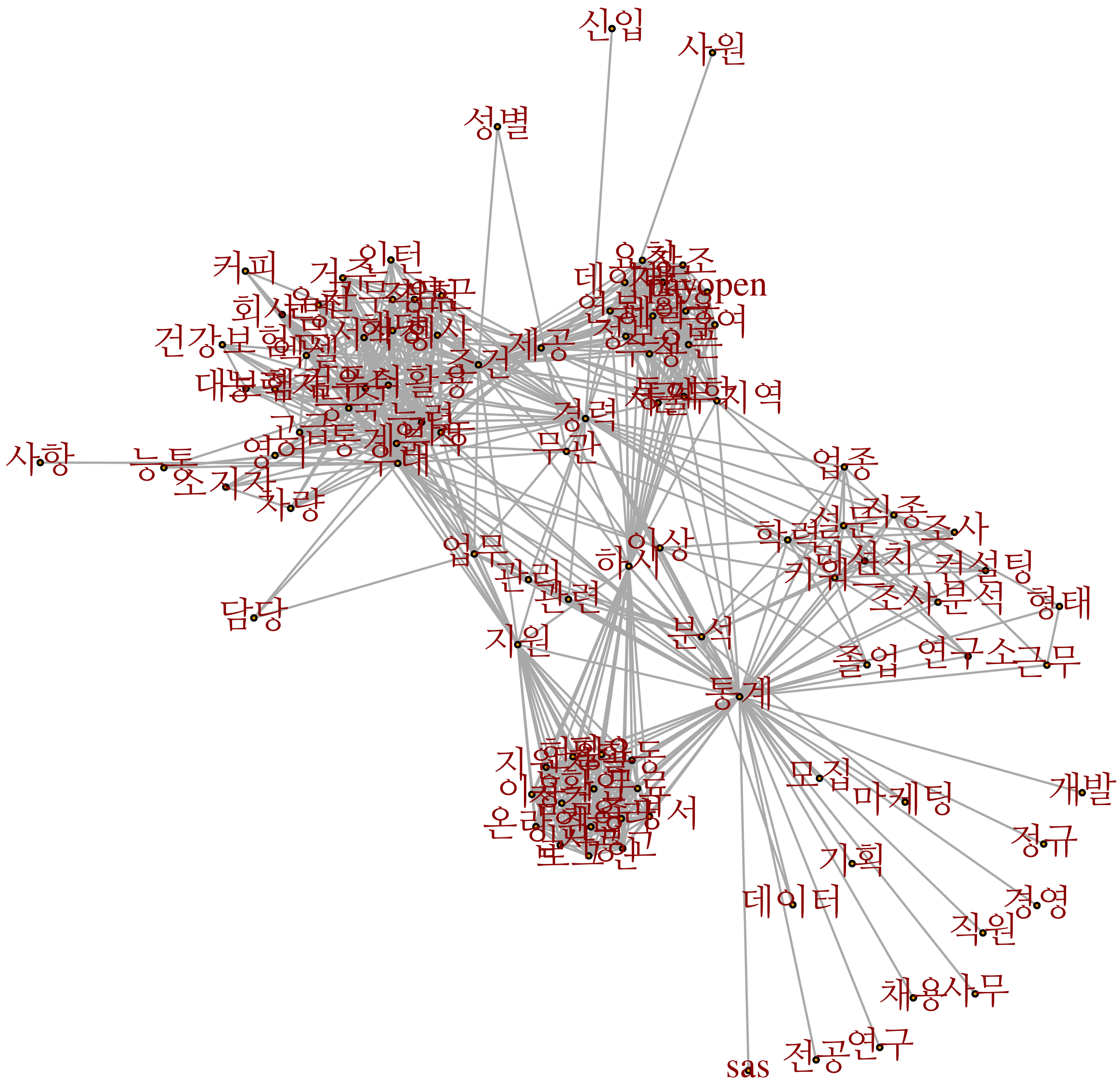
5.10 Keywords association network

- Positive correlation >= 0.7

```
diag(x) <- 0
x[(x < 50)] <- 0
didx <- apply(x, 1, sum) > 0 & apply(x, 2, sum) > 0
sum(didx)
```

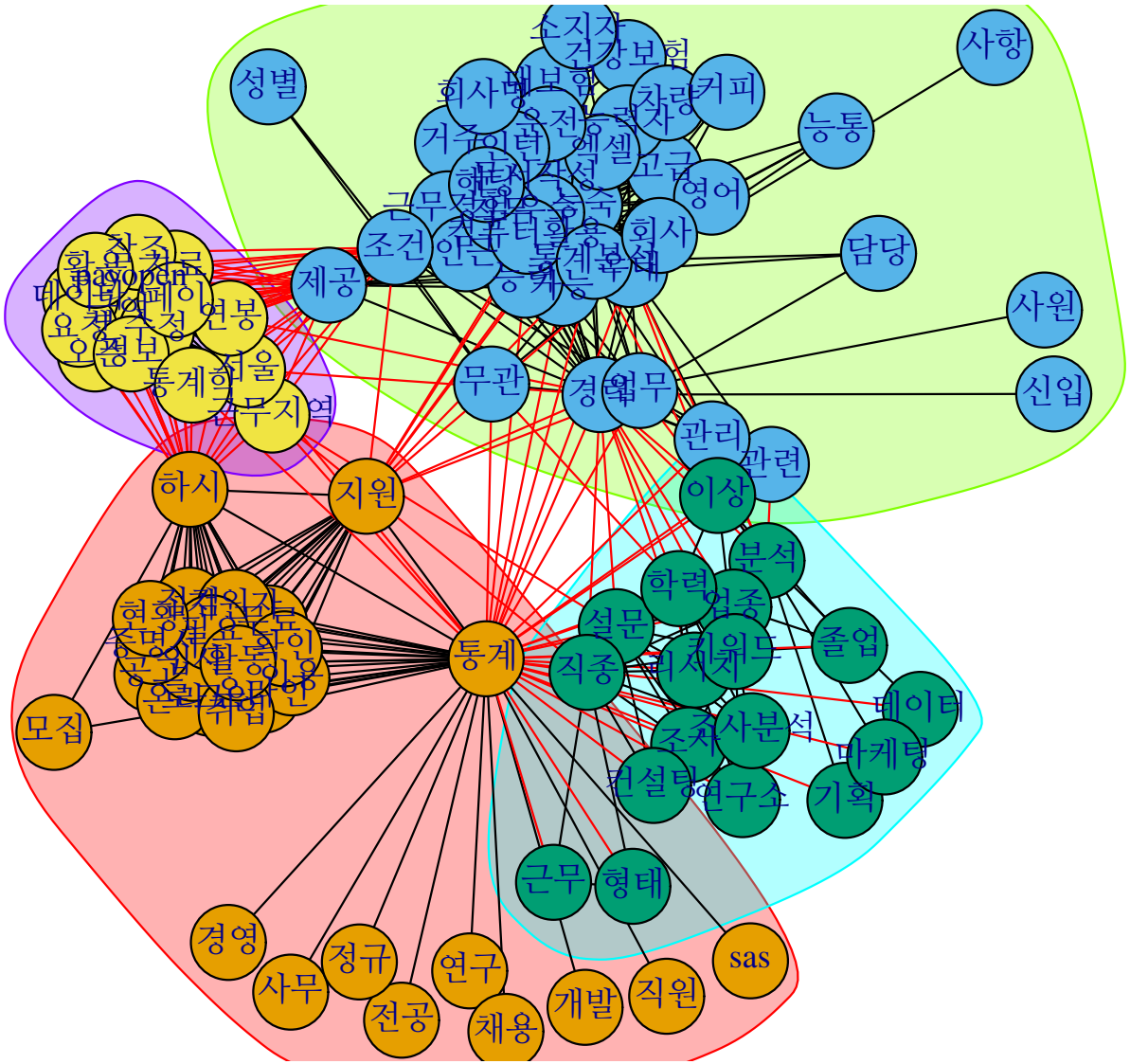
```
## [1] 100
```

```r
cc <- x[didx, didx]
# graph 객체로 변환
g <- graph.adjacency(cc,  mode="undirected")
g <- simplify(g) # remove loops
# set labels and degrees of vertices
V(g)$label <- rownames(cc) #단어
set.seed(10)
plot(g, vertex.size=1, vertex.label.color="darkred")
```


신입
사원
성별
커피
인턴
제공
조건
지역
경력
무관
사항
능통
영어
우대
차량
업무
관리
관련
담당
지원
이상
분석
업종
학력
조사
컨설팅
형태
조사분석
졸업
연구소
근무
통계
개발
모집
마케팅
정규
기획
데이터
경영
직원
채용사무
전공연구
sas

5.11 커뮤니티 그래프(walktrap.community)

```
wc <- walktrap.community(g)
plot(wc, g, margin=-0.1)
```


성별
사항
능통
영어
조건
제공
담당
사원
신입
무관
경력
관리
관련
하시
지원
이상
분석
학력
업종
설문
통계
직종
졸업
모집
근무
형태
경영
사무
정규
연구
전공
채용
개발
직원
sas

```
# 커뮤니티 멤버십
table(wc$mem)
```

```
##
##  1  2  3  4
## 28 39 18 15
```

5.12 멤버십별로 서브그래프를 드로잉

```
    sub <- wc$members==5;
    plot(subgraph(g, sub))
}
```

6 참고문헌

- 정우영, 한승희 (2008). 구인광고 분석을 통한 국내 정보전문직의 취업동향 분석. 한국정보관리학회 학술대회 논문집, 157-164.
- 권영옥 (2013). 빅데이터를 활용한 맞춤형 교육 서비스 활성화 방안연구. 지능정보연구, 19(2), 87-100.
- 오은주 (2016). NCS 기반 교육과정 적용 연구.

- 최윤희, 송민, 김기영 (2014). 마이닝기법을 활용한 사서커뮤니티 내 채용정보 및 인력수급현황 분석. 한국정보관리학회 학술대회 논문집, 143-147.
- Clauset, Newman, Moore (2004). Finding community structure in very large networks, Physical review E, 70(6), 066111. http://kr.indeed.com/